Day 1: Fundamental data and statistical literacy

Consuming statistics

Simon Munzert
Hertie School

Table of contents

Making sense of descriptive statistics

Descriptive vs. inferential statistics

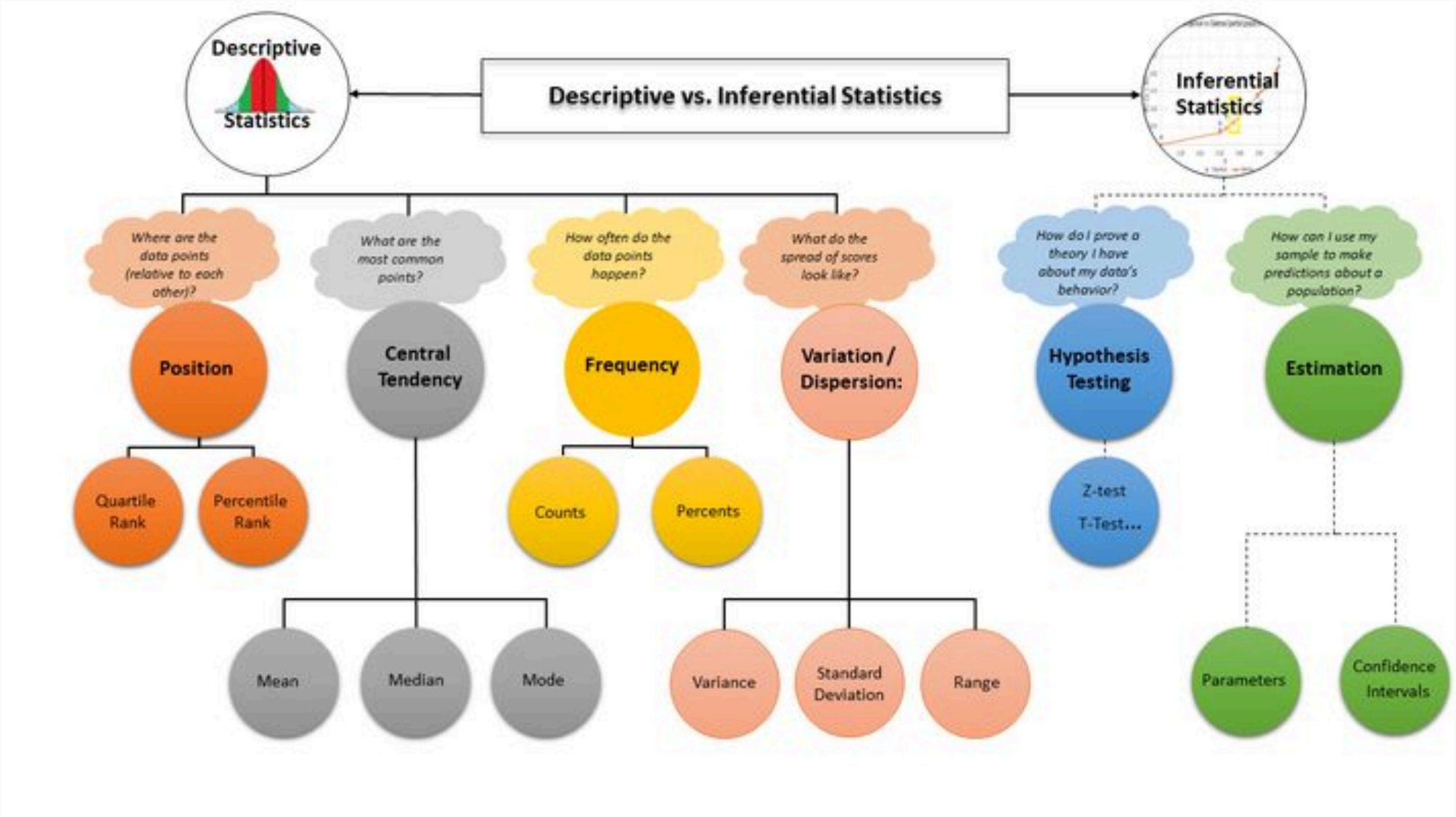
Descriptive statistics

- Summarize and describe characteristics of a sample or population
- Can be communicated numerically and visually
- Different scales (levels of measurement) require different descriptive statistics
- Good description can be challenging if data collection or measurement is complex

Inferential statistics

- Make inferences about a population based on a sample
- Can be inferences about means, proportions, relationships, etc.
- Can be communicated numerically and visually
- Good description is the basis for good inference

Descriptive vs. inferential statistics



Measures of central tendency

Three popular measures of central tendency

- **(Arithmetic) Mean**: The **average** of all values in a dataset
- **Median**: The **middle** value of a dataset
- **Mode**: The **most frequent** value in a dataset

Why "central tendency"? Describes the tendency of quantitative data to cluster around some central value.

Try it out

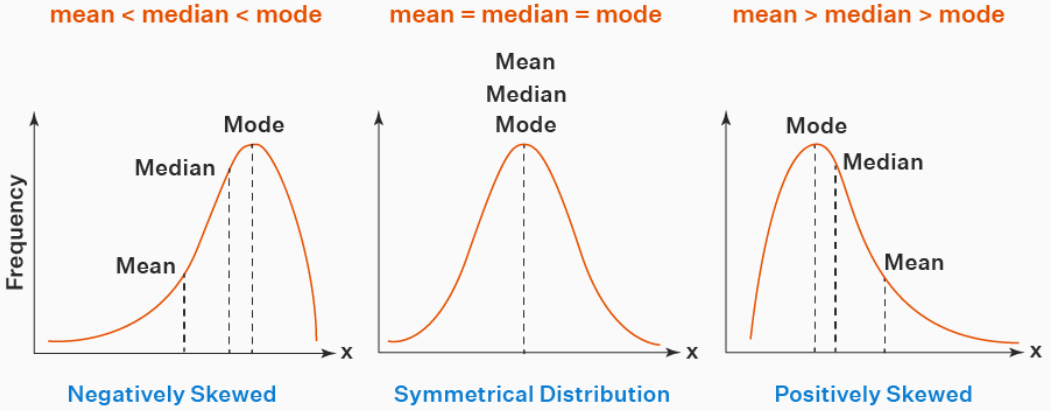
Identify the mode, median, and mean of the following values:

8, 2, 4, 2, 18, 6, 2

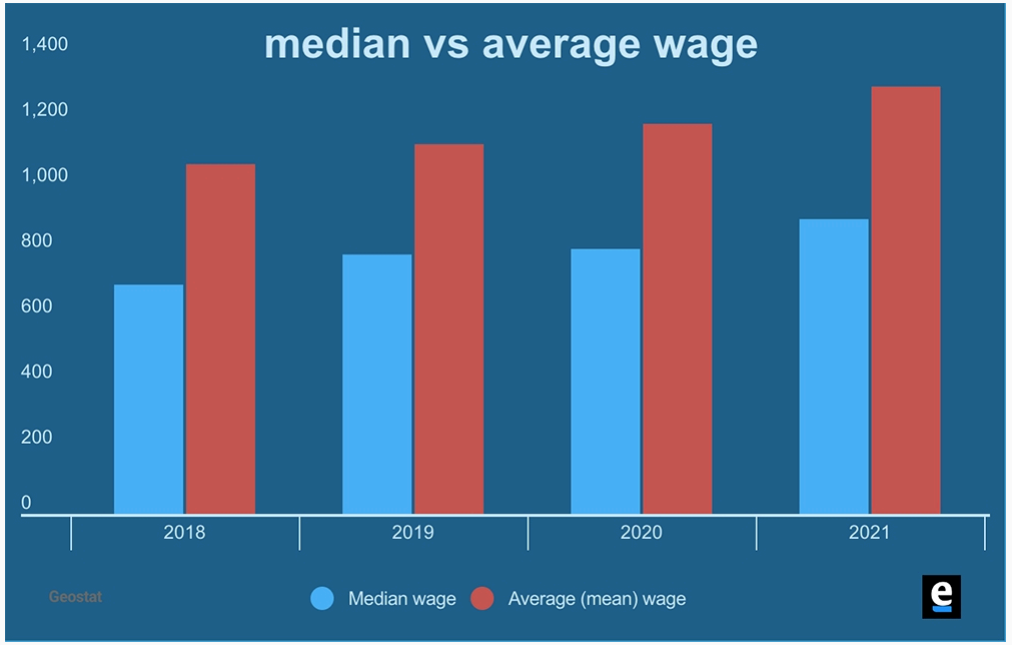
Which one to use?

- **Mean**: Sensitive to outliers, but with an intuitive meaning
- **Median**: Robust to outliers, but a bit less intuitive
- **Mode**: Useful for categorical data, but can be misleading for continuous data

Skewed distributions

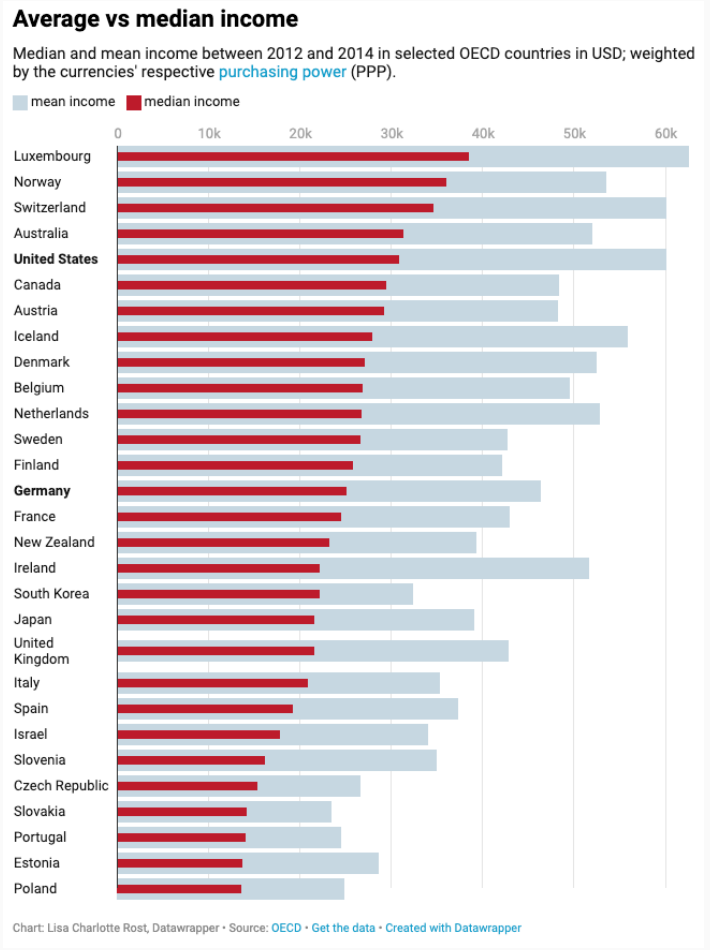


Measures of central tendency: examples



(Measured in Lari)

`Source` eurasianet, 2022



`Source` Lisa Muth, Datawrapper

Measures of variation

Why do we need measures of variation?

- Central tendency alone does not tell the whole story
- "How spread out/stretched are our data?"

Three popular measures of variation

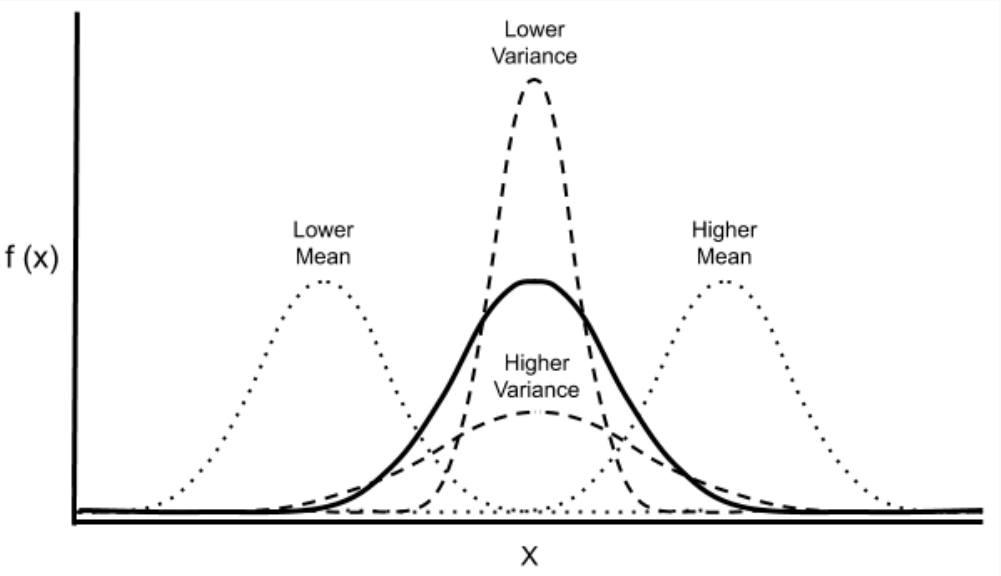
- **Range**: The difference between the highest and lowest value in a dataset
- **Variance**: The average of the squared differences from the mean
- **Standard deviation**: The square root of the variance

Formula to calculate the variance: $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

See here for interactive intuition.

Why is this substantively important?

- Most people are, in fact, not "average". Variation can be a source of insight into underlying processes
- Key measure for downstream statistics, e.g., the standard error as an estimate of sampling variability (uncertainty of an estimate)



Puzzle 1

How can all of the following be true?[1]

1. 80% of the 100 most prominent Georgian TikTokers are male.
2. On average, female Georgian TikTokers have 500 followers whereas males only 300.
3. There's an approximately equal number of male and female Georgian TikTokers.

[1] "True" in the sense of "theoretically true". The numbers are all made up.

Puzzle 2

How can all of the following be true?[1]

1. At a Georgian university, the acceptance rate to each of four departments is higher for females than for males.
2. Aggregated over the departments, the acceptance rate is higher for males.

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	792	34%	417	33%	375	35%
D	714	6%	373	6%	341	7%
Total	**3024**	**39%**	**2175**	**47%**	**849**	**31%**

Legend:

greater percentage of successful applicants than the other gender

greater number of applicants than the other gender

bold - the two 'most applied for' departments for each gender

[1]Same disclaimer as in previous example.

Simpson's paradox

Paradox explained

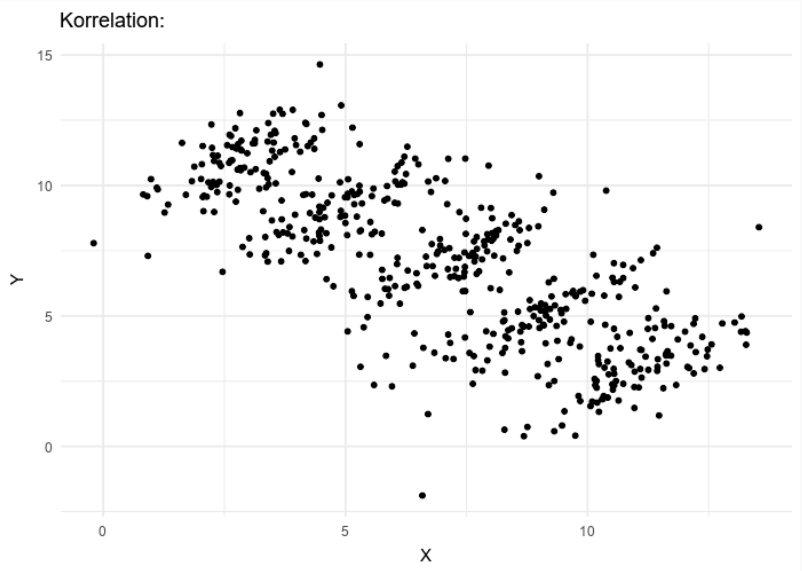
- Some departments (C+D) were more competitive than others, and women were applying more to those.
- Overall acceptance rates and within-department acceptance rates have different baselines!
- It is not really a "paradox", but a confounding issue: When we adjust/control for the grouping variable, the relationship between the variables changes.

The phenomenon, generalized

- A trend appears in different groups of data but disappears or reverses when these groups are combined.
- This can also arise in correlations (positive vs. negative correlation within vs. across groups).

Relevance for policy-making

- Analysis of patterns at different levels (e.g., regional vs. federal, schools vs. school districts)
- If within-group patterns are not accounted for, policy conclusions might be misleading.



`Source` Wikipedia, "Simpson's paradox"

Making sense of probability

Probabilities

Hertie School

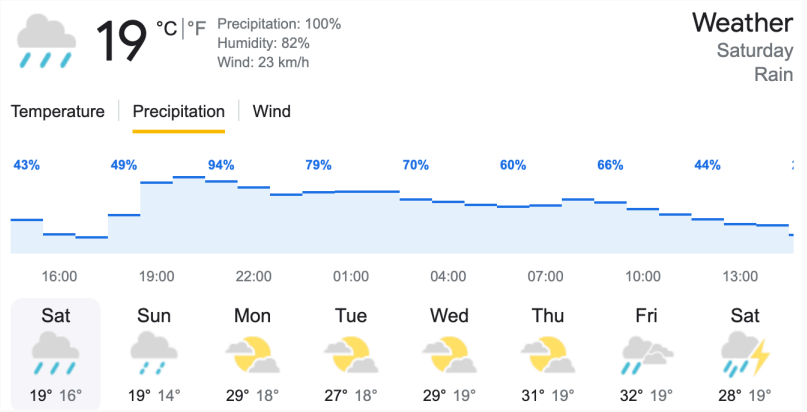
What are probabilities?[1]

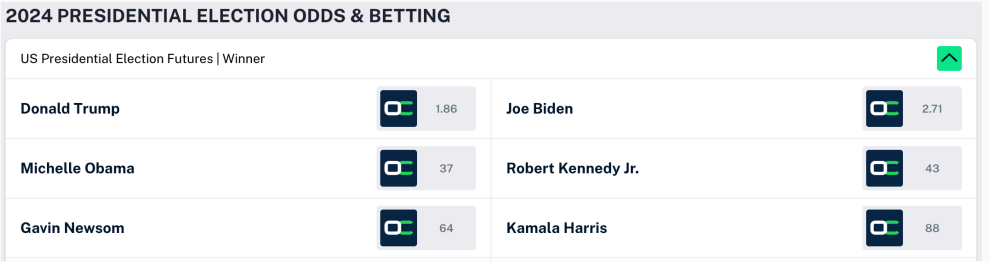
- Probabilities quantify the likelihood of an event occurring
- Probabilities are between 0 and 1 (or 0 and 100%)
- Probabilities can be communicated verbally or numerically

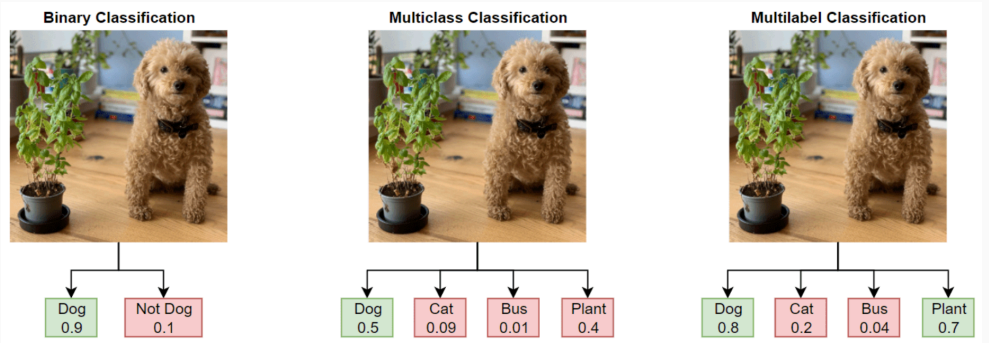
Relevance of probabilities for policy-making

Probabilities are ...

- ... at the core of risk assessment and decision-making
- ... used to quantify uncertainty
- ... used to assess the effectiveness of policies







[1]See here for a nice primer to probability and simulation.

Marginal, conditional, and joint probabilities

Marginal probability

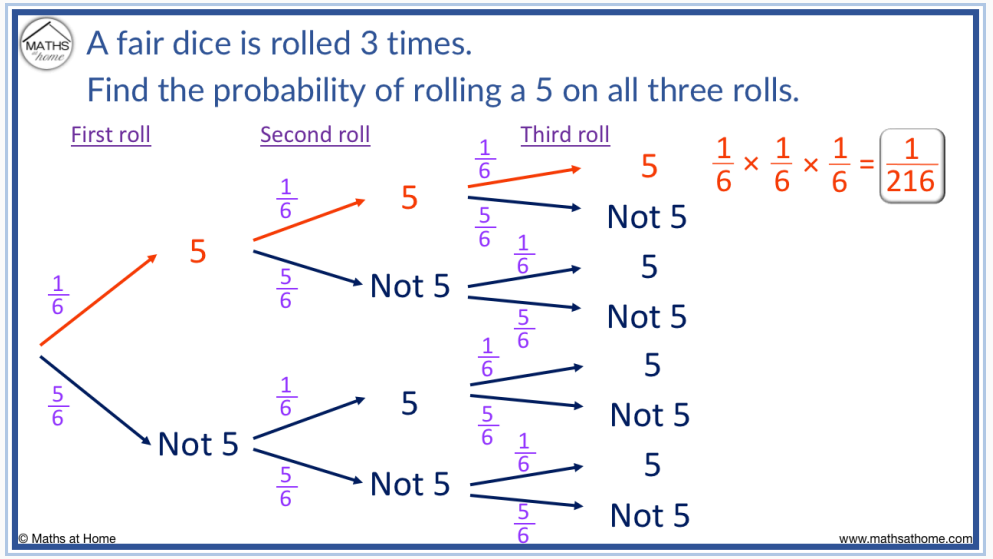
- The probability of an event occurring: $p(A)$
- Unconditional probability - it is not conditioned on another event
- 🎲 example: $p(\text{rolling a } 5) = 1/6$

Conditional probability

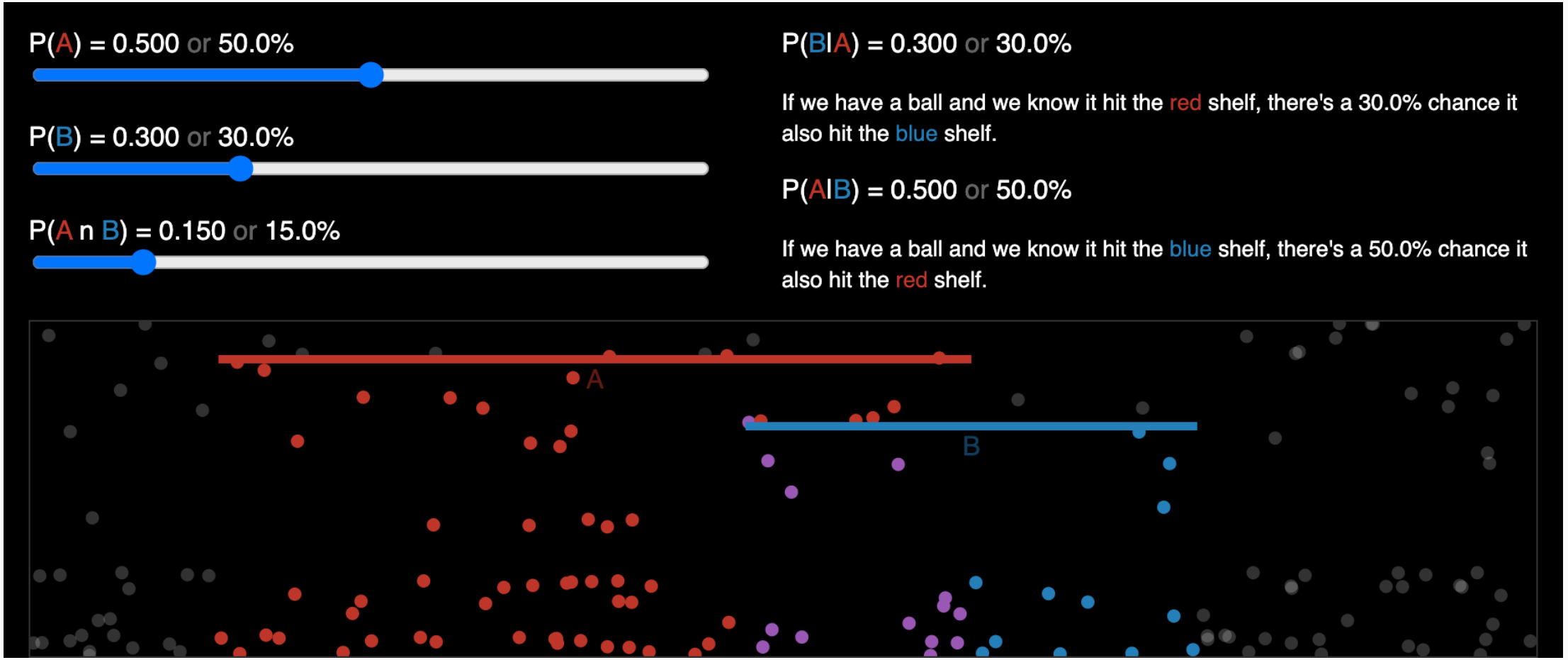
- The probability of event A occurring, given that event B occurs: $p(A|B)$
- Important: The marginal probability of B does not matter here!
- 🎲 example: $p(\text{rolling a } 5 | \text{rolling an odd number}) = 1/3$

Joint probability

- The probability of event A and event B occurring: $p(\text{A and B}) = p(A \cap B)$
- 🎲 example: $p(\text{rolling a } 5 \text{ and an even number}) = 0$



Conditional probabilities



`Source` Victor Powell, setosa.io (check out for interactive simulation)

Puzzle

How can all of the following be true?

1. A vaccine is highly effective in protecting against a disease.
2. Most people who get the disease have been vaccinated.

`Source` Hakan Nural, Unsplash

Puzzle

How can all of the following be true?

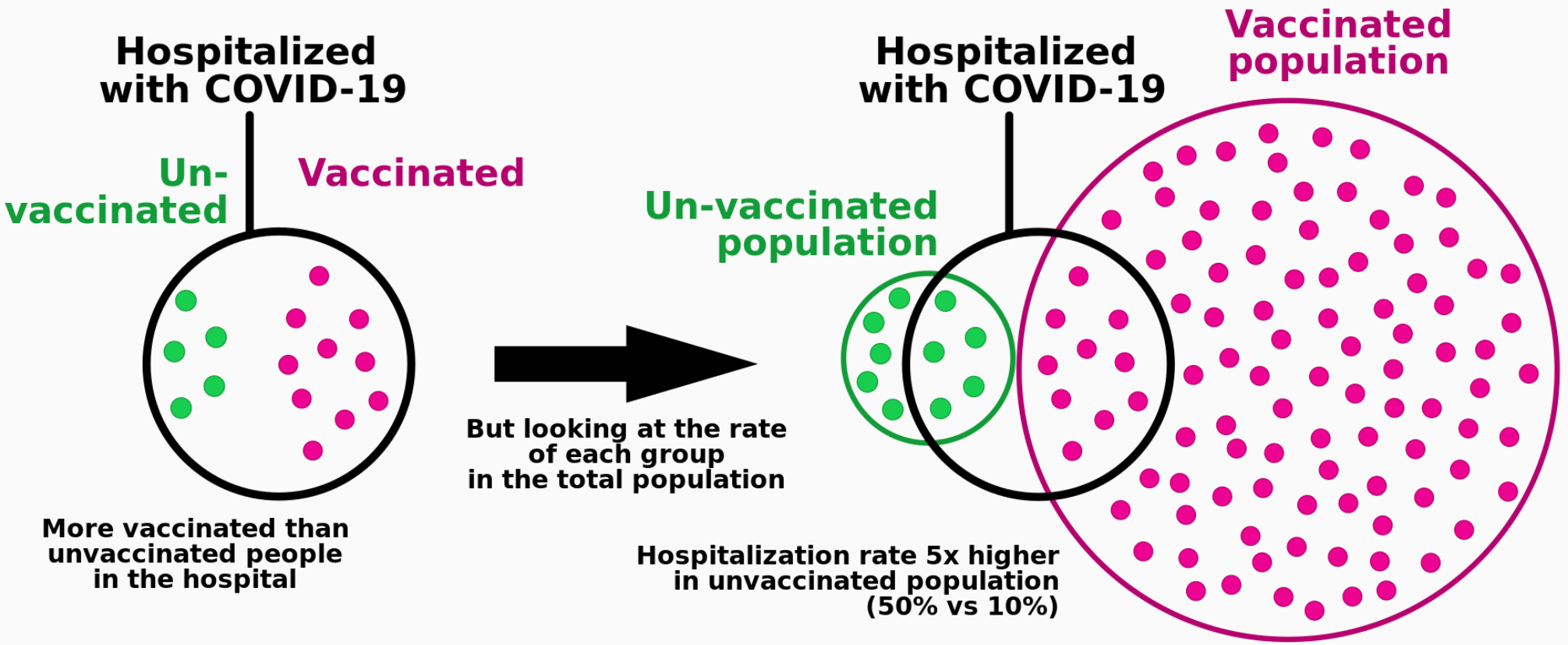
1. A vaccine is highly effective in protecting against a disease.
2. Most people who get the disease have been vaccinated.

Base rate fallacy

- This is a classic case of the "base rate fallacy" or "prosecutor's fallacy".
- If the vaccination rate P(vaccinated) is high in the population, there's just much more opportunity for vaccinated people to get hospitalized than for unvaccinated people.

`Source` Hakan Nural, Unsplash

The base rate fallacy illustrated



`Source` Marc Rumilly

Relevance of the base rate fallacy for policy

Policies for low-probability events

- Some policies are designed to prevent low-probability events that are extremely costly if they happen.
- Examples: Terrorist attacks, war, natural disasters
- Predicting such events is inherently difficult
- Yet, AI-supported detection systems claim high detection rates. But even with very high accuracy, the amount of false positives generated can be prohibitively high

Guidance

- Always consider the **base rate** when interpreting probabilities
- Be cautious when interpreting **conditional probabilities** without considering the base rate

Example: terrorist identification

- In a city of 1m inhabitants, there are 100 terrorists and 999,900 non-terrorists: $p(\text{terrorist}) = 0.0001$
- Surveillance system based on facial recognition software with two failure rates of 1%:
  1. False negative rate: $p(\text{no alarm}|\text{terrorist}) = 0.01$
  2. False positive rate: $p(\text{alarm}|\text{no terrorist}) = 0.01$

What does this mean when we get an alarm?[1]

$$p(\text{terrorist}|\text{alarm}) = \frac{p(\text{alarm}|\text{terrorist})p(\text{terrorist})}{p(\text{alarm})} = \frac{0.99*0.0001}{0.01} = 0.01$$

[1] Getting $p(\text{alarm}) = p(\text{alarm}|\text{terrorist}) * p(\text{terrorist}) + p(\text{alarm}|\text{no terrorist}) * p(\text{no terrorist}) = p(\text{terrorist}|\text{alarm}) = 0.99 * 0.0001 + 0.01 * 0.9999 = 0.01$

Communicating probabilities with verbal expressions

Hertie School

JCOM

Variability in the interpretation of probability phrases used in Dutch news articles — a risk for miscommunication

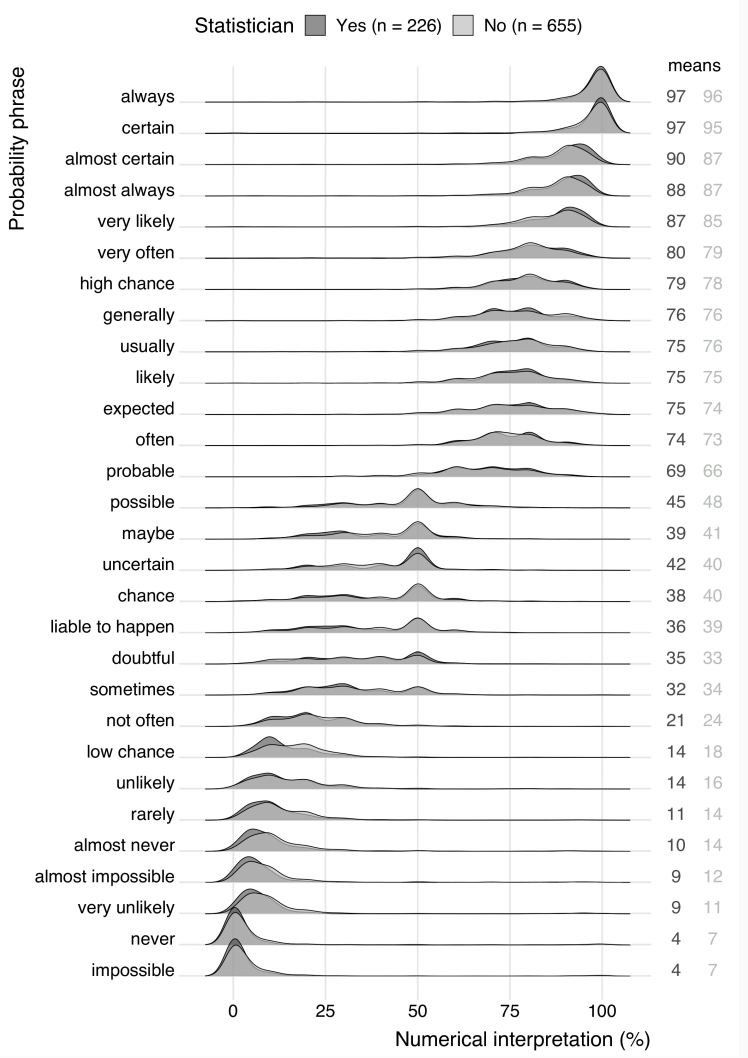
Sanne Willems, Casper Albers and Ionica Smeets

Abstract

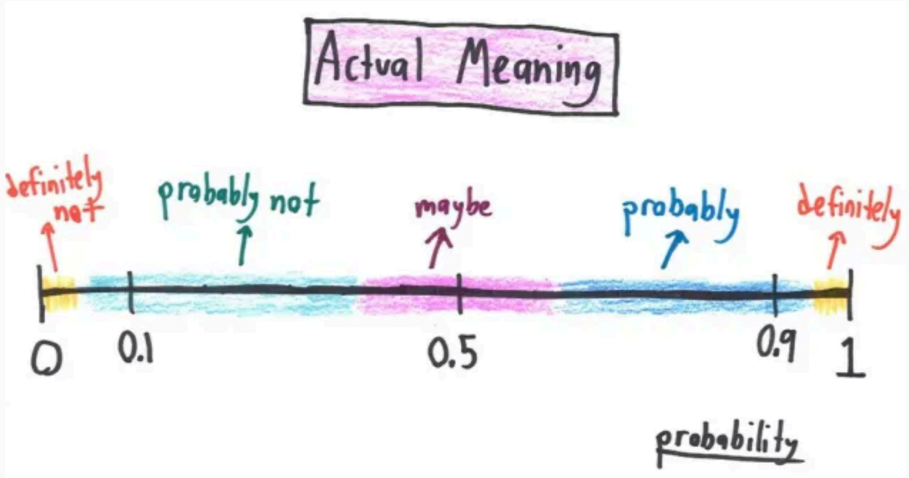
Verbal probability phrases are often used in science communication to express estimated risks in words instead of numbers. In this study we look at how laypeople and statisticians interpret Dutch probability phrases that are regularly used in news articles. We found that there is a large variability in interpretations, even if the phrases are given in a neutral context. Also, statisticians do not agree on the interpretation of the phrases. We conclude that science communicators should be careful in using verbal probability expressions.

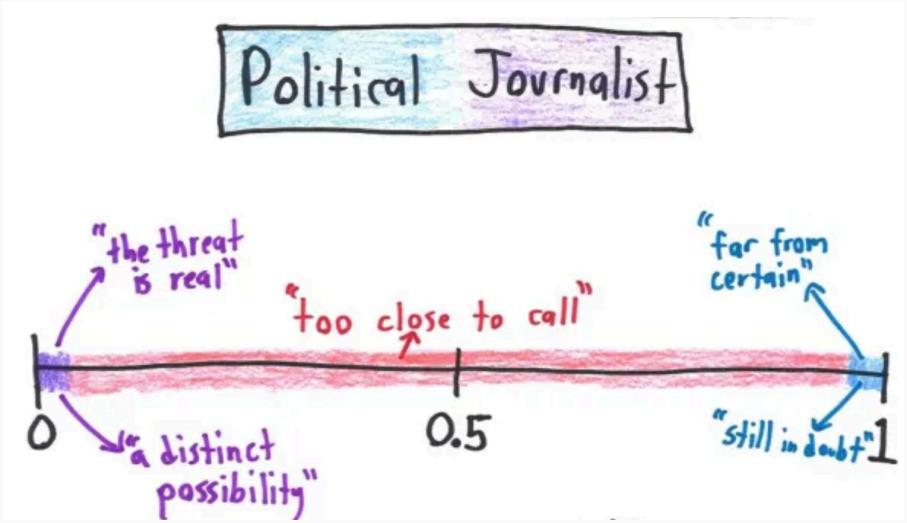
Keywords Risk communication; Science and media; Science writing

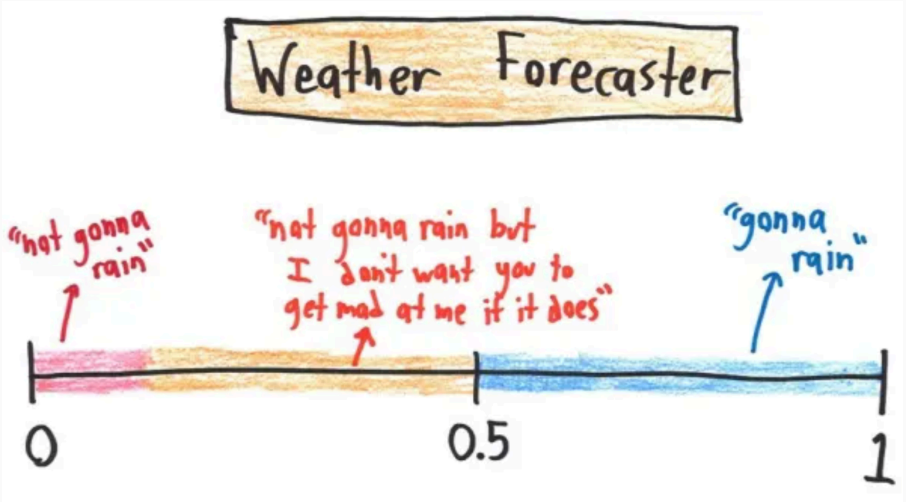
`Source` Willems et al. 2020



What does probability mean for different professions?







Source Ben Orlin

Making sense of statistical effects

Statistical effects and effect sizes

What are statistical effects?

- Statistical effect quantify a difference or relationship between variables

Example questions about effects

- What is the marginal effect of education on income?
- How much does the probability of voting increase with age?

Statistical effect $\neq$ causal effect!

- Statistical effects are about statistical relationships between variables, not about causal relationships.
- For instance, just because your model tells you that an additional 100$ per month income are associated with 1 additional year of education, you would not conclude that income causally increases education

Effect sizes

- Effect sizes are quantitative measures of the **strength of a relationship**.
- Effect sizes express the magnitude of a difference or relationship in a standardized way.

Examples

- A (unstandardized or standardized) **group mean difference**
- The **correlation coefficient** is an effect size for the relationship between two continuous variables (see later!), e.g., $r = 0.1 \rightarrow$ weak, $r = 0.5 \rightarrow$ moderate, $r = 0.9 \rightarrow$ strong
- The **regression coefficient** expresses the predicted marginal change in an outcome relative to a unit change of the predictor (potentially conditionally on other covariates)

Example: regression effects

	Hourly wage
Education	0.505***
	(0.051)
Female	−2.275***
	(0.279)
Nonwhite	−0.119
	(0.460)
Intercept	0.650
	(0.681)
N	526
R^2	0.259
Adjusted R^2	0.255

Some advice when consuming effect sizes

Some questions to ask yourself

1. What does the effect size mean **substantively**? E.g., what does an "effect of 0.87" mean?

	Prominence	Influence
Senate	0.906***	1.483***
	(0.060)	(0.067)
Sessions served	0.163***	0.292***
	(0.016)	(0.017)
Party (Independent)	0.701*	1.059**
	(0.368)	(0.412)
Party (Republican)	0.035	−0.080
	(0.047)	(0.052)
Office: Governor	0.266*	0.450**
	(0.158)	(0.177)
Office: Lt. Governor	−0.031	0.089
	(0.257)	(0.288)
Office: US Secretary	0.551**	0.372
	(0.262)	(0.294)
Position: House Speaker	1.896***	2.670***
	(0.385)	(0.431)
Position: Majority/Minority Leader	0.185	0.711**
	(0.308)	(0.345)
Position: Whip	0.231	0.848***
	(0.233)	(0.261)
Position: Deputy Whip	0.698***	0.462*
	(0.234)	(0.262)
Position: Party Chairman	−0.115	−0.255
	(0.215)	(0.241)
(Intercept)	1.648***	1.527***
	(0.050)	(0.057)
N	492	492
R-squared	0.493	0.694
Adj. R-squared	0.481	0.687
Residual Std. Error (df = 479)	0.505	0.565
F Statistic (df = 12; 479)	38.890***	90.715***

***p < .01; **p < .05; *p < .1

Some advice when consuming effect sizes

Some questions to ask yourself

1. What does the effect size mean **substantively**? E.g., what does an "effect of 0.87" mean?

2. Is the effect size **plausible**? How does it compare to your intuition and other effects in the literature?

Political Science Research and Methods (2020), page 1 of 7
doi:10.1017/psrm.2019.63

RESEARCH NOTE

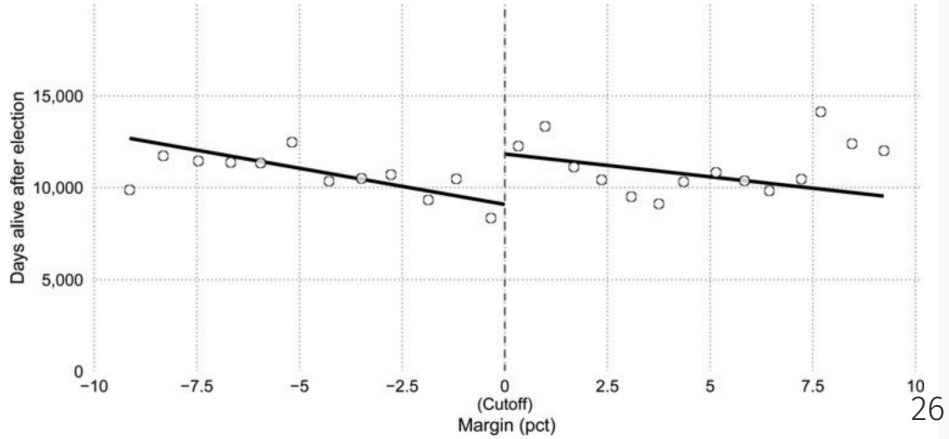
Longevity returns to political office

Sebastian Barfort[1], Robert Klemmensen[2] and Erik Gahner Larsen[3]*

[1]Independent Researcher, Copenhagen, Denmark, [2]Political Science, University of Southern Denmark, Odense, Denmark and [3]School of Politics and International Relations, University of Kent, Canterbury, Kent, United Kingdom of Great Britain and Northern Ireland
*Corresponding author. E-mail: E.G.Larsen@kent.ac.uk

(Received 18 April 2019; revised 19 September 2019; accepted 13 November 2019)

Abstract
Does political office cause worse or better longevity prospects? Two perspectives in the literature offer contradicting answers. First, increased income, social status, and political connections obtained through holding office can increase longevity. Second, increased stress and working hours associated with holding office can have detrimental effects on longevity. To provide causal evidence, we exploit a regression discontinuity design with unique data on the longevity of candidates for US gubernatorial office. The results show that politicians winning a close election live 5–10 years longer than candidates who lose.



Some advice when consuming effect sizes

Some questions to ask yourself

1. What does the effect size mean **substantively**? E.g., what does an "effect of 0.87" mean?

2. Is the effect size **plausible**? How does it compare to your intuition and other effects in the literature?

3. How **precisely** is the effect estimated?

Making sense of statistical significance

"Statistical significance" everywhere

Attractive names sustain increased vegetable intake in schools

Brian Wansink [a,*], David R. Just [b], Collin R. Payne [c], Matthew Z. Klinger [d]

[a] Department of Applied Economics and Management at Cornell University, 15 Warren Hall, Ithaca, NY 14853-7801, USA
[b] Department of Applied Economics and Management at Cornell University, 16 Warren Hall, Ithaca, NY 14853-7801, USA
[c] New Mexico State University, College of Business, MSC 5280, PO Box 30001, Las Cruces, NM 88003-8001, USA
[d] Half Hollow Hills High School East, 50 Vanderbilt Parkway, Dix Hills, NY 11746, USA

A B S T R A C T

Objective: This study will determine if the selective use of attractive names can be a sustainable, scalable means to increase the selection of vegetables in school lunchrooms.

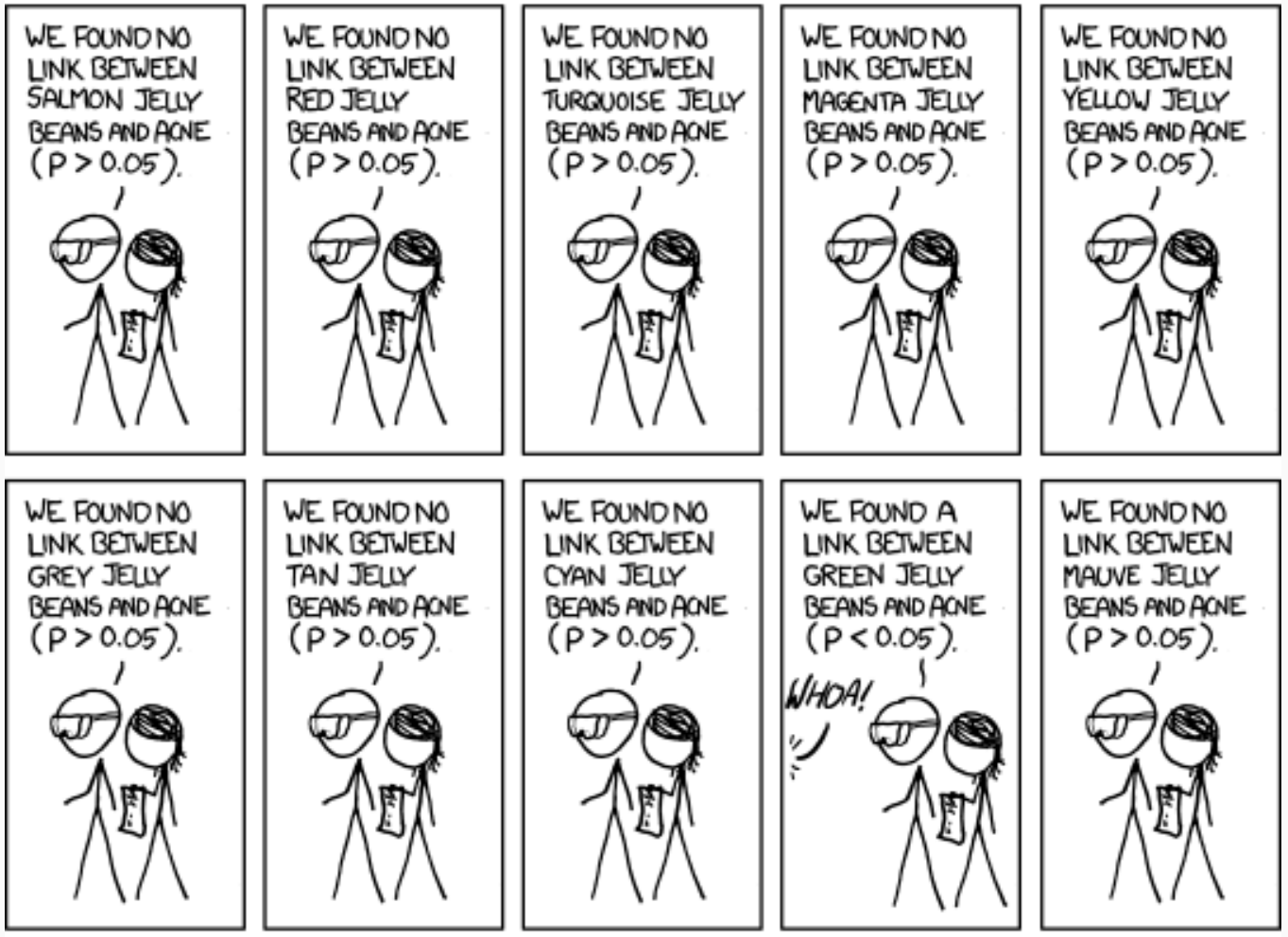
Methods: Study 1 paired an attractive name with carrots in five elementary schools (n = 147) and measured selection and consumption over a week compared to controls. Study 2 tracked food sales of vegetables in two elementary schools (n = 1017) that were systematically attractively named or not named over a two-month period. Both studies were conducted in New York in 2011.

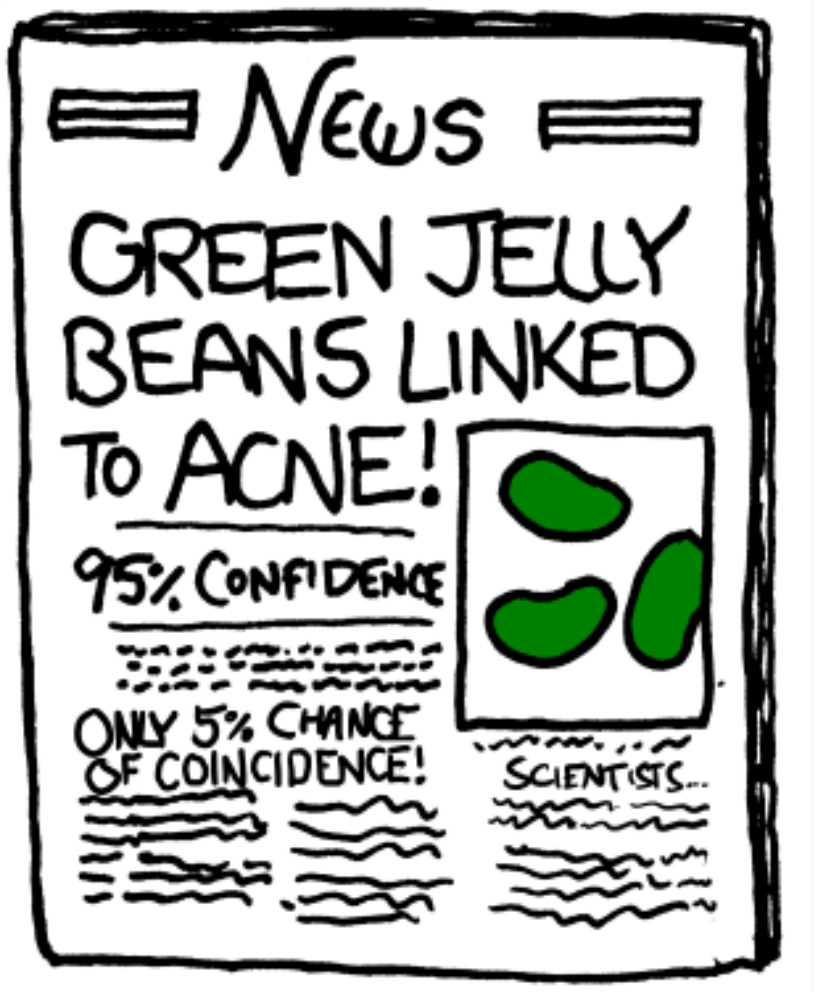
Results: Study 1 found that elementary students ate twice the percentage of their carrots if attractively named as "X-ray Vision Carrots," than if un-named or generically named as the "Food of the Day." Study 2 found that elementary school students were 16% more likely to persistently choose more hot vegetable dishes ($p<0.001$) when they were given fun or attractive names.

Discussion: Attractive names effectively and persistently increased healthy food consumption in elementary schools. The scalability of this is underscored by the success of Study 2, which was implemented and executed for negligible cost by a high school student volunteer.

`Source` Wansink et al., Retraction Watch

"Statistical significance" everywhere

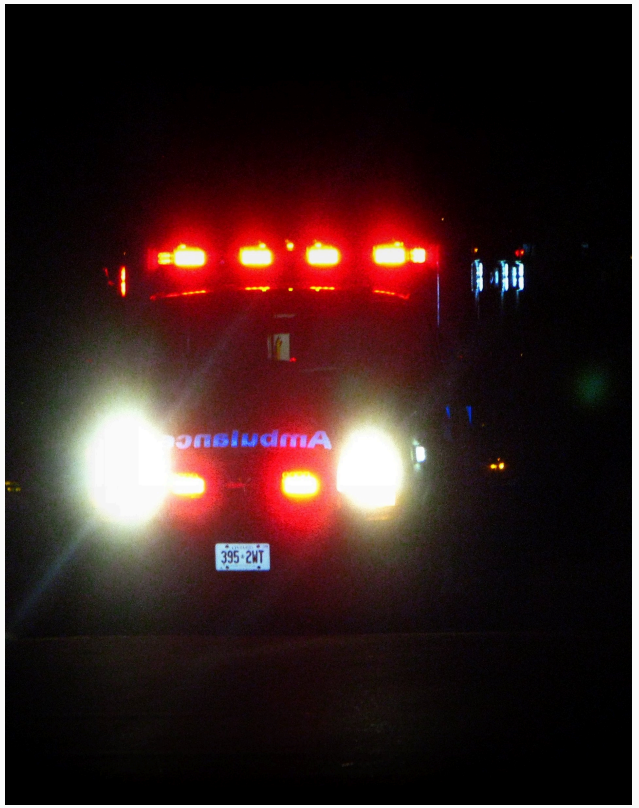




Statistical error types

A life and death example of statistical errors

You are a paramedic and you approach the scene of a car accident. One victim is laying motionless on the road and you must assess whether the victim is dead or alive, and the victim will be treated accordingly. Based on this information, **which error results in the most costly mistake?**



Hypotheses

- **Null hypothesis**: The victim is alive.
- **Alternative hypothesis**: The victim is not alive.

`Source` jeffalltogether, StackExchange.com

Statistical error types

Hypotheses

- **Null hypothesis**: The victim is alive.
- **Alternative hypothesis**: The victim is not alive.

Error types

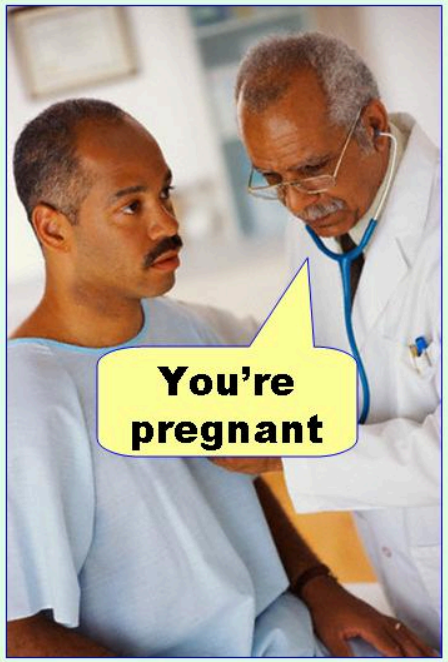
- **Type I error**: You reject the null when the null is actually true. ("false positive")
- **Type II error**: You fail to reject the null when the null is actually false. ("false negative")

Costs

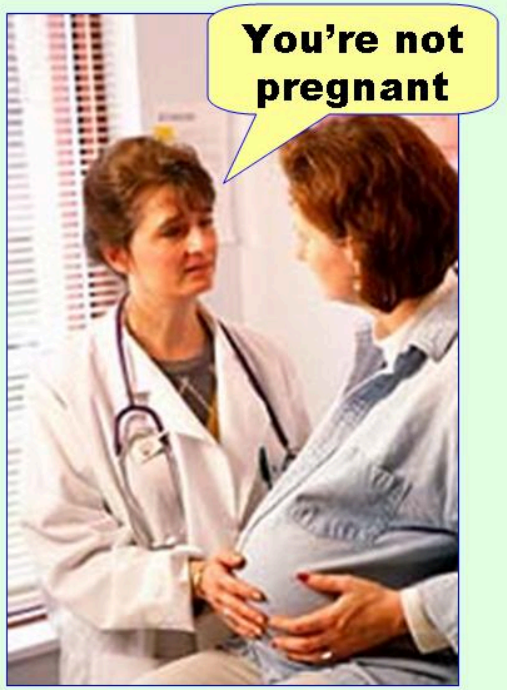
- **Type I error**: You declare the victim dead when they are actually alive. They do not receive an ambulance to the hospital for a life saving medical treatment. → **Extremely costly mistake**
- **Type II error**: You declare the victim alive when they are actually dead. You erroneously send a dead person to the hospital in an ambulance → **Not that costly mistake**

Error types illustrated

Type I error
(false positive)



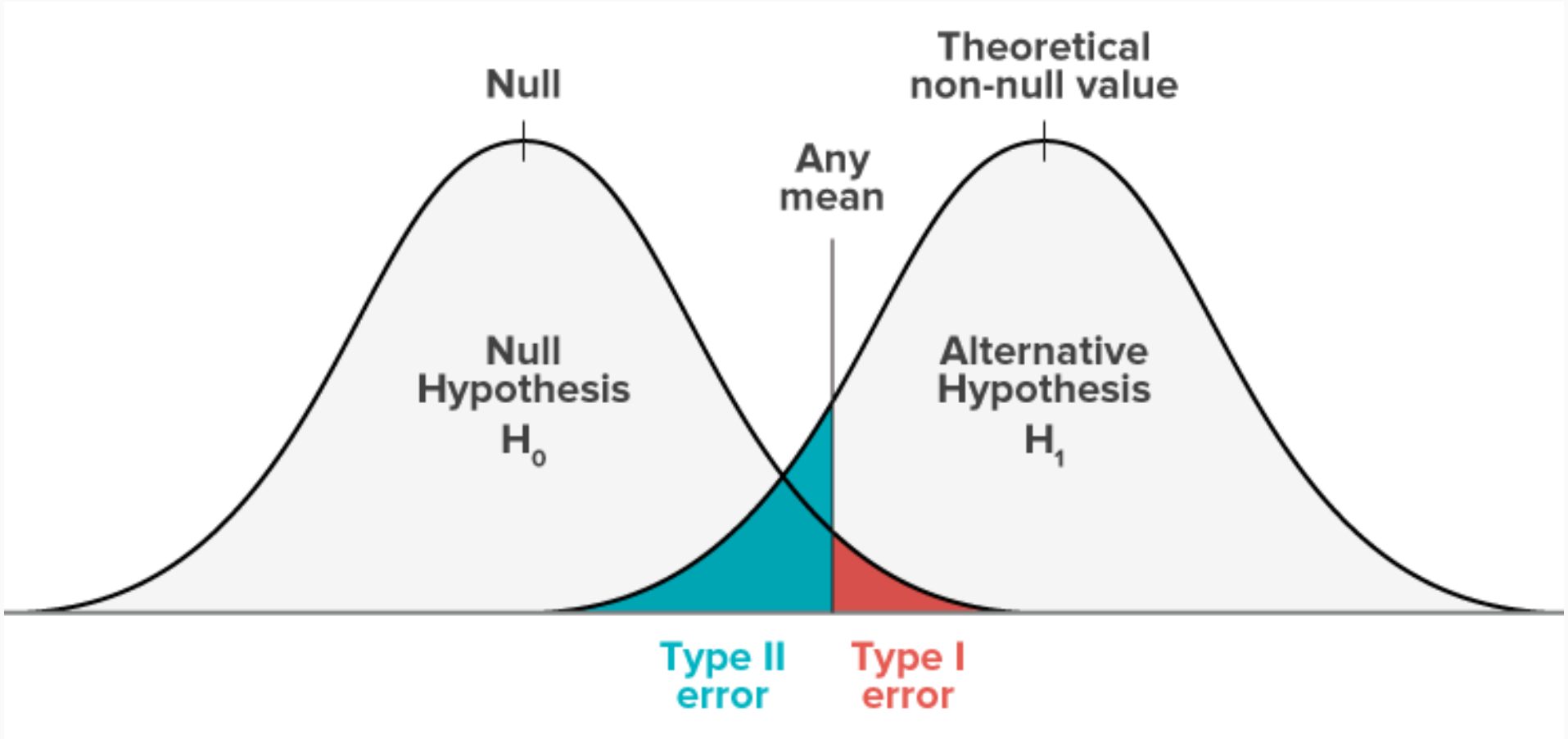
Type II error
(false negative)



Error types in hypothesis testing

		The Truth (Based on Entire Population)	
		Nothing Is There (H_0 Is True)	Something Is There (H_0 Is False)
Your Conclusion (Based on Your Sample)	I Don't See Anything (Nonsignificant)	Right!	Wrong (Type II Error)
	I See Something (Significant)	Wrong (Type I Error)	Right!

Error types in hypothesis testing



Assessing statistical significance

Statistical significance vs. practical significance

- They are not the same.
- Statistical significance is about the probability of observing the data given the null hypothesis.
- Practical significance is about the real-world importance of the result.

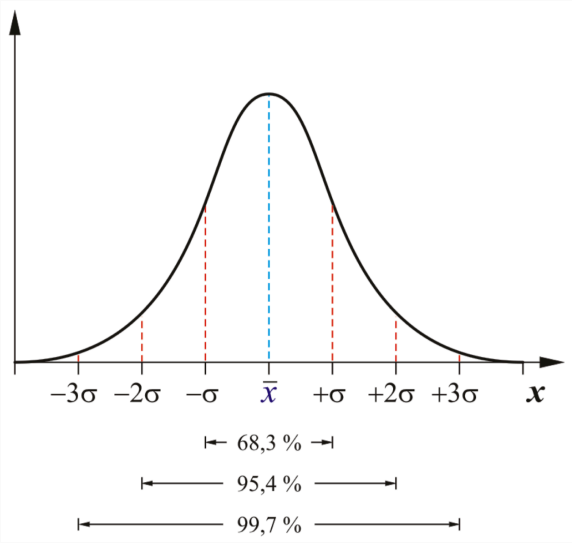
From hypothesis testing to statistical significance

Three-step approach:

1. Formulate null and alternative hypotheses.
2. Calculate a test statistic. For instance, effect size in a regression divided by standard error.
3. Compare the test statistic to a critical value; calculate a p-value.

The p-value

- The p-value is the probability of observing a result at least as extreme as the observed result if the null hypothesis were true.
- The p-value is compared to a threshold (e.g., 0.05) to decide whether to reject the null hypothesis.
- Importantly, the p-value is not the probability that the null hypothesis is true or false!



Eyeballing statistical significance

Hertie School

	Prominence	Influence
Senate	0.906***	1.483***
	(0.060)	(0.067)
Sessions served	0.163***	0.292***
	(0.016)	(0.017)
Party (Independent)	0.701*	1.059**
	(0.368)	(0.412)
Party (Republican)	0.035	−0.080
	(0.047)	(0.052)
Office: Governor	0.266*	0.450**
	(0.158)	(0.177)
Office: Lt. Governor	−0.031	0.089
	(0.257)	(0.288)
Office: US Secretary	0.551**	0.372
	(0.262)	(0.294)
Position: House Speaker	1.896***	2.670***
	(0.385)	(0.431)
Position: Majority/Minority Leader	0.185	0.711**
	(0.308)	(0.345)
Position: Whip	0.231	0.848***
	(0.233)	(0.261)
Position: Deputy Whip	0.698***	0.462*
	(0.234)	(0.262)
Position: Party Chairman	−0.115	−0.255
	(0.215)	(0.241)
(Intercept)	1.648***	1.527***
	(0.050)	(0.057)
N	492	492
R-squared	0.493	0.694
Adj. R-squared	0.481	0.687
Residual Std. Error (df = 479)	0.505	0.565
F Statistic (df = 12; 479)	38.890***	90.715***

***p < .01; **p < .05; *p < .1

Table 1. Summary statistics

Variable	South (1)	North (2)	Difference in means (3)	Adjusted difference in means (4)	*P* value (5)
Panel 1: Air pollution exposure at China's Disease Surveillance Points					
TSPs, μg/m^3	354.7	551.6	196.8***	199.5***	<0.001/0.002
SO_2, μg/m^3	91.2	94.5	3.4	−3.1	0.812/0.903
NO_x, μg/m^3	37.9	50.2	12.3***	−4.3	<0.001/0.468
Panel 2: Climate at the Disease Surveillance Points					
Heating degree days	2,876	6,220	3,344***	482	<0.001/0.262
Cooling degree days	2,050	1,141	−910***	−183	<0.001/0.371
Panel 3: Demographic features of China's Disease Surveillance Points					
Years of education	7.23	7.57	0.34	−0.65	0.187/0.171
Share in manufacturing	0.14	0.11	−0.03	−0.15***	0.202/0.002
Share minority	0.11	0.05	−0.05	0.04	0.132/0.443
Share urban	0.42	0.42	0.00	−0.20*	0.999/0.088
Share tap water	0.50	0.51	0.02	−0.32**	0.821/0.035
Rural, poor	0.21	0.23	0.01	−0.33*	0.879/0.09
Rural, average income	0.34	0.33	0.00	0.24	0.979/0.308
Rural, high income	0.21	0.19	−0.02	0.27	0.772/0.141
Urban site	0.24	0.25	0.01	−0.19	0.859/0.241
Predicted life expectancy	74.0	75.5	1.54***	−0.24	<0.001/0.811
Actual life expectancy	74.0	75.5	1.55	−5.04**	0.158/0.044

The sample (n = 125) is restricted to DSP locations within 150 km of an air quality monitoring station. TSP (μg/m^3) in the years 1981–2000 before the DSP period is used to calculate city-specific averages. Degree days are the deviation of each day's average temperature from 65°F, averaged over the years 1981–2000 before the DSP period. The results in column (4) are adjusted for a cubic in degrees of latitude north of the Huai River boundary. Predicted life expectancy is calculated by OLS using all of the demographic and meteorological covariates shown. All results are weighted by the population at the DSP location. One DSP location is excluded due to invalid mortality data. *Significant at 10%, **significant at 5%, ***significant at 1%. Sources: China Disease Surveillance Points (1991–2000), *China Environment Yearbook* (1981–2000), and World Meteorological Association (1980–2000).

Eyeballing statistical significance

Hertie School

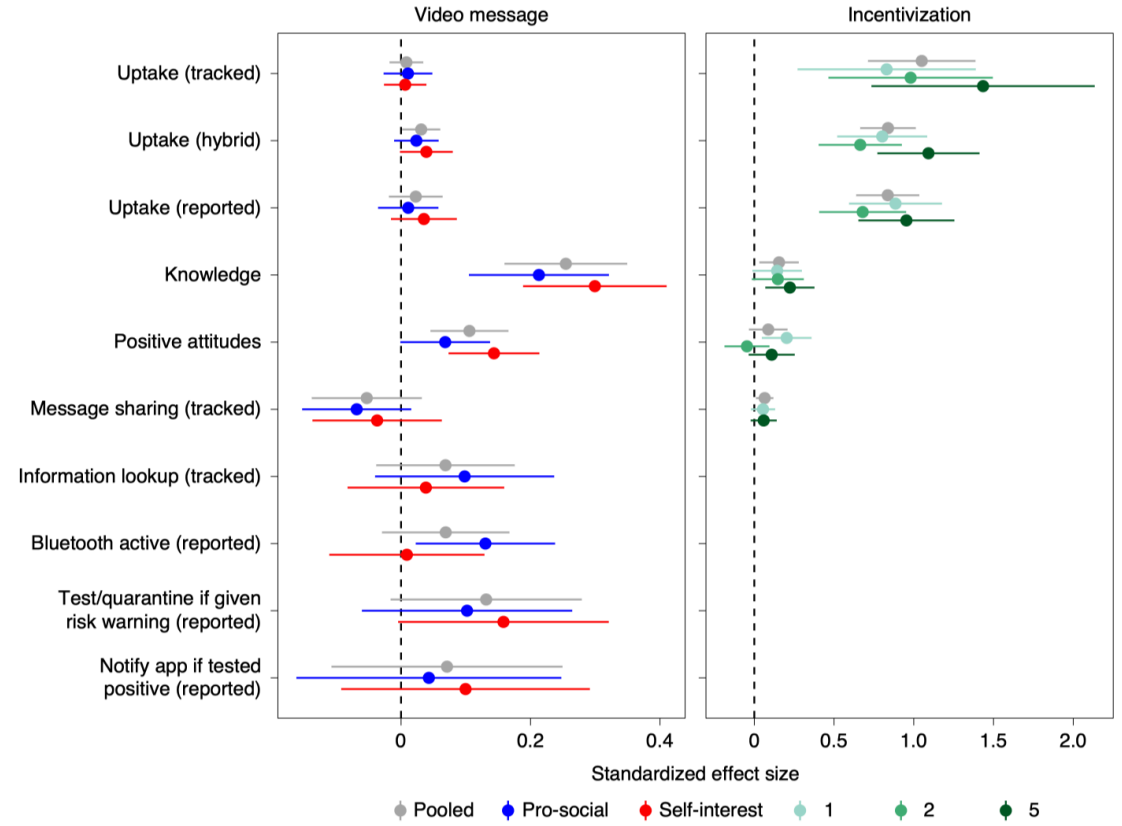


Fig. 3 | Effect of message and incentive treatments on uptake, knowledge, attitudes and behaviour. Each plot shows standardized ITT estimates with 95% CIs from fully saturated ordinary least squares regression models fit using the pre-registered LASSO covariate selection procedure. The video message sample comprises $n = 2{,}044$, 1,356 and 1,337 respondents for estimation of the pooled, pro-social and self-interest treatment effects, respectively. The incentive sample comprises $n = 1{,}015$, 513, 516 and 494 respondents for estimation of the pooled, €1, €2 and €5 treatment effects, respectively.

Controversies around statistical significance[1]

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

The ASA's Statement on *p*-Values: Context, Process, and Purpose

Six principles

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]

Received: 9 April 2016 / Accepted: 9 April 2016 / Published online: 21 May 2016

Abstract Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific literature. In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the *P* values they produce) can lead to small *P* values even if the declared test hypothesis is correct, and can lead to large *P* values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of *P* values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.

Editor's note This article has been published online as supplementary material with an article of Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process and purpose. The American Statistician 2016.

[1]See also here for a nice primer to this controversy.

Consuming statistics: lessons learned

Lies, Damned Lies and Statistics

- Policy debates almost inevitable also revolve around statistics
- Strategic incentive to tilt evidence in favor
- Statistical pitfalls: Not everything that sounds logical is statistically sound
- A fundamental understanding of basic concepts of statistics is key to make you a critical consumer of statistical information
- Some popular fallacies and errors occur again and again → learn to spot them!